A Comprehensive Analysis on Machine Learning based Methods for Lung Cancer Level Classification

Shayli Farshchiha[a], Salman Asoudeh[b], Maryam Shavali Kuhshuri[c], Mehrshad Eisaei[d] and Mohamadreza Azadi[e]

[a] Isfahan University of Technology, Department of Electrical and Computer Engineering, Faculty of Engineering, Isfahan, Iran.

[b] Faculty of Computer, College of Technical and Engineering, University of Velayat, Iran.

[c] University of Isfahan Department of Electrical Engineering, Faculty of Engineering, University of Isfahan, Isfahan, Iran.

[d] University of Mazandaran, Mazandaran, Iran.

[e] Department of Computer Science,Shamsipour technical college Tehran, Iran.

Abstract

Lung cancer remains a significant global health concern, necessitating accurate and efficient diagnostic methodologies. This study undertakes a comprehensive exploration into the utilization of machine learning (ML) techniques for the precise classification of lung cancer levels. A meticulous examination is conducted to mitigate overfitting challenges while optimizing model efficacy, incorporating strategies such as monitoring minimum child weight and fine-tuning learning rates. Diverse ML models, encompassing XGBoost (XGB), LGBM, Adaboost, Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), CatBoost, and k-Nearest Neighbor(k-NN), are meticulously employed and rigorously evaluated. Furthermore, the study scrutinizes the intricate relationships between features and targets through the lens of Deep Neural Networks (DNN), elucidating their potential for complex pattern recognition. The findings underscore the efficacy of various ML models in accurately stratifying lung cancer levels. Despite the intricacies of DNN architectures, conventional ML models, notably XGBoost, LGBM, and Logistic Regression, demonstrate exceptional performance. Comprehensive comparative metrics including accuracy, precision, recall, and F-1 score underscore the superiority of select models in the realm of lung cancer prediction.

Keywords: Lung cancer prediction, Machine learning, Overfitting, Model performance, Deep Neural Networks.

1. Introduction

Lung cancer stands as a pervasive global health challenge, holding the grim distinction of being the foremost cause of cancer-related deaths and one of the most commonly diagnosed cancers worldwide. Each year, approximately 2.20 million individuals receive the devastating diagnosis of lung cancer, with a troubling 75% facing mortality within five years. The intricate inter-tumor heterogeneity and the development of drug resistance present formidable hurdles in effective cancer treatment. Nevertheless,

recent advancements in cancer research technologies have ushered in an era of collaborative projects, yielding extensive databases of clinical, imaging, and sequencing data. These repositories empower researchers to delve comprehensively into the intricate patterns of lung cancer, spanning from diagnosis and treatment to clinical outcomes. Notably, the emergence of -omics analyses, including genomics, transcriptomics, proteomics, and metabolomics, has significantly expanded our investigative toolkit.

However, the utilization of diverse and high-dimensional data types for clinical applications necessitates substantial time and expertise. This underscores the increasing significance of machine learning (ML) models, which autonomously discern underlying data patterns to aid physicians in decision-making processes. Despite their potential, there remains a reluctance among practitioners and researchers to invest resources in decision-making processes involving costly datasets. ML offers a promising solution by streamlining decision-making, ensuring precision, cost-effectiveness, and reproducibility, irrespective of geographical or personnel constraints.

In our study, we endeavor to leverage ML models to develop precise prediction tools for classifying lung cancer types based on essential protein attributes. Through the integration of feature selection, tree induction, and clustering techniques, we seek to augment the predictive capabilities of these models and contribute to the advancement of lung cancer diagnosis and treatment. Additionally, we address the challenges surrounding clinical and in-situ patient monitoring, aiming to provide cost-effective and time-efficient solutions.

ML, as a subset of artificial intelligence (AI), has long served as an invaluable aid in cancer phenotyping and therapy. Our research employs a diverse array of ML models, encompassing both conventional and modern algorithms, to offer comprehensive insights for patients, researchers, and domain experts alike. Furthermore, we meticulously address commonly overlooked hyperparameters in ML models, such as min child weights and learning rates, elucidating their profound impact on accuracy errors through detailed analysis. By pushing the boundaries of ML techniques in lung cancer diagnosis and management, our study aims to improve patient outcomes and streamline decision-making processes in clinical settings.

2. Related Works

Machine learning has emerged as a cornerstone of modern healthcare, offering unprecedented opportunities to enhance diagnostic accuracy and prognostic capabilities in diseases such as lung cancer. By harnessing the power of machine learning algorithms, researchers are not only advancing our understanding of disease mechanisms but also revolutionizing clinical practice through personalized treatment strategies and predictive modeling. In the context of lung cancer, predicting post-operative life expectancy is of paramount importance for treatment planning and patient counseling. The utilization of predictive data mining algorithms, including Decision Tree, Naive Bayes, and Artificial Neural Network, enables clinicians to tailor treatment regimens based on individual patient profiles [1]. Moreover, the incorporation of stratified 10-fold cross-validation analyses ensures the robustness and generalizability of these predictive models, thereby enhancing their clinical utility.

Furthermore, the application of classification algorithms for brain tumor detection offers valuable insights applicable to lung cancer diagnosis. By evaluating the performance of various classifiers against volumetric and locational features, researchers gain crucial insights into the discriminatory power of

these algorithms [2]. Such insights not only facilitate early detection and intervention but also enable clinicians to delineate tumor characteristics and tailor treatment strategies accordingly.

In the realm of lung cancer detection, the remarkable accuracy demonstrated by classifiers such as Support Vector Machine underscores the transformative potential of machine learning in clinical practice [3]. By leveraging diverse algorithms, including KNN, SVM, NN, and Logistic Regression, clinicians can make more informed decisions regarding diagnosis, staging, and treatment selection. Moreover, the integration of machine learning-based diagnostic tools into clinical workflows holds promise for improving patient outcomes and optimizing resource allocation within healthcare systems.

Additionally, the exploration of segmentation algorithms such as Naïve Bayes and Hidden Markov Model further enhances the diagnostic precision of lung tumor detection [4]. By delineating tumor boundaries and characterizing tissue morphology, these algorithms enable clinicians to identify subtle abnormalities and differentiate between benign and malignant lesions [5–8]. Furthermore, the development of algorithmic flowcharts for brain tumor detection underscores the interdisciplinary nature of research in oncological diagnostics, with implications for lung cancer diagnosis and treatment [9].

Overall, the integration of machine learning algorithms into clinical practice represents a paradigm shift in oncological diagnostics and personalized medicine. By harnessing the power of data-driven insights, clinicians can improve patient outcomes, optimize resource allocation, and ultimately, advance our collective understanding of lung cancer and its management. As research in this field continues to evolve, the promise of machine learning in transforming healthcare delivery remains unparalleled.

3. Materials and Methods

Predicting lung cancer using machine learning poses significant challenges due to the need to collect and organize datasets from diverse sources worldwide. This is essential for ensuring balanced datasets representative of various regions and scenarios. To address this, data for detecting and classifying lung cancer levels are collected from public datasets shared by reputable organizations such as the World Health Organization (WHO), Kaggle, and Google datasets. By aggregating data from multiple sources, we aim to create an approximately balanced dataset.

3.1 Dataset

The first step involves collecting data from different public datasets shared by reputable organizations such as WHO, Kaggle, and Google datasets. The collected data include various attributes such as patient ID, age, gender, and several environmental and lifestyle factors known to influence lung cancer risk. The target variable is the level of lung cancer, categorized into three stages: low, medium, and high.

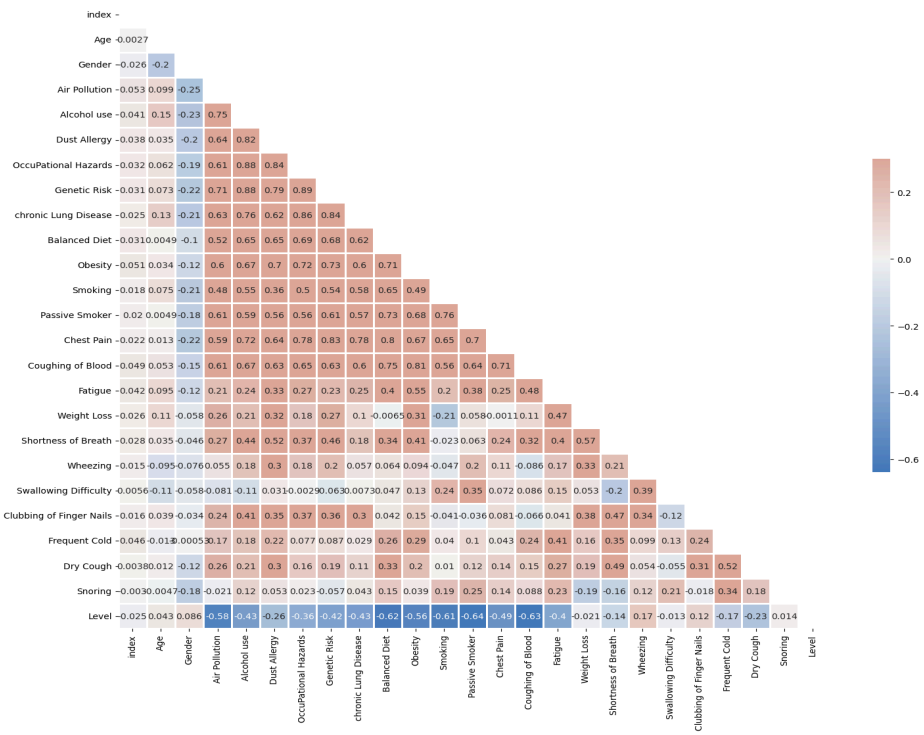

Figure 1. Lung Cancer Correlation.

3.2 Preprocessing

Once collected, the datasets undergo cleaning and preprocessing to establish meaningful relationships between the attributes. Features are examined for correlations to gain insights into their impact on the target variable. Correlation plots are generated to visualize the relationships between features. Features with correlations above a certain threshold (e.g., 0.5) are considered for combination to create new columns in the dataset. This process enhances the effectiveness of the features in predicting lung cancer levels.

3.3 Correlation Analysis

Correlation plots are used to examine the relationships between features and the target variable. Features with high positive correlations (e.g., above 0.5) are merged to create new columns, improving predictive accuracy. Conversely, features with low or negative correlations are analyzed to determine their impact on the target variable.

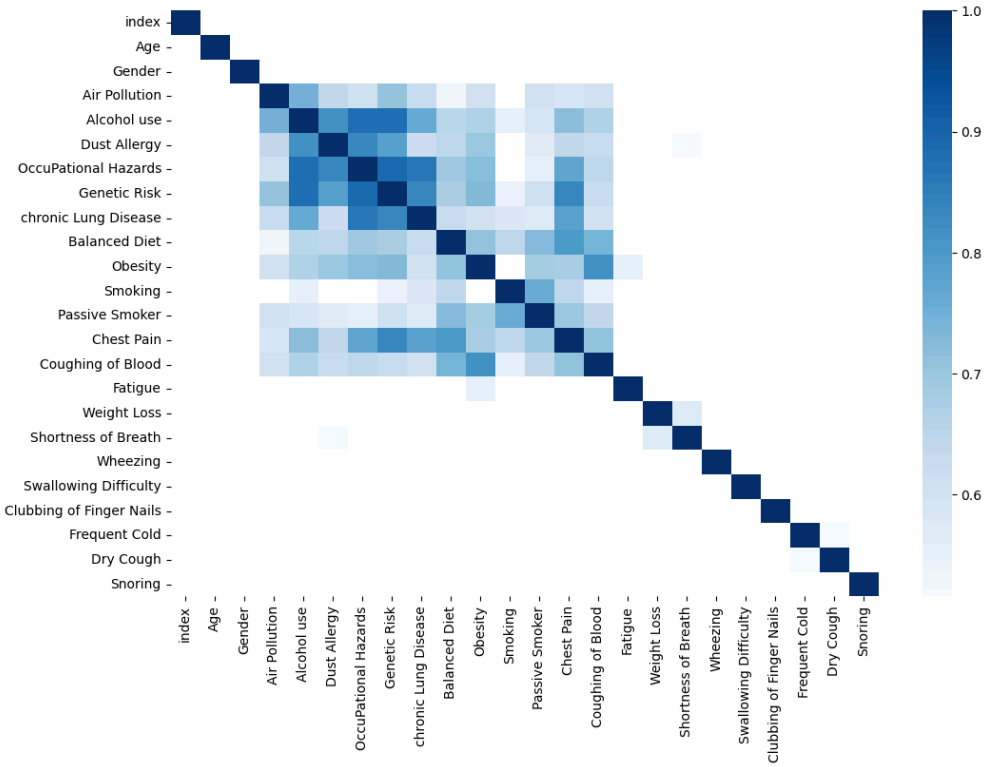

Figure 2. Lung Cancer Correlation More than 0.5.

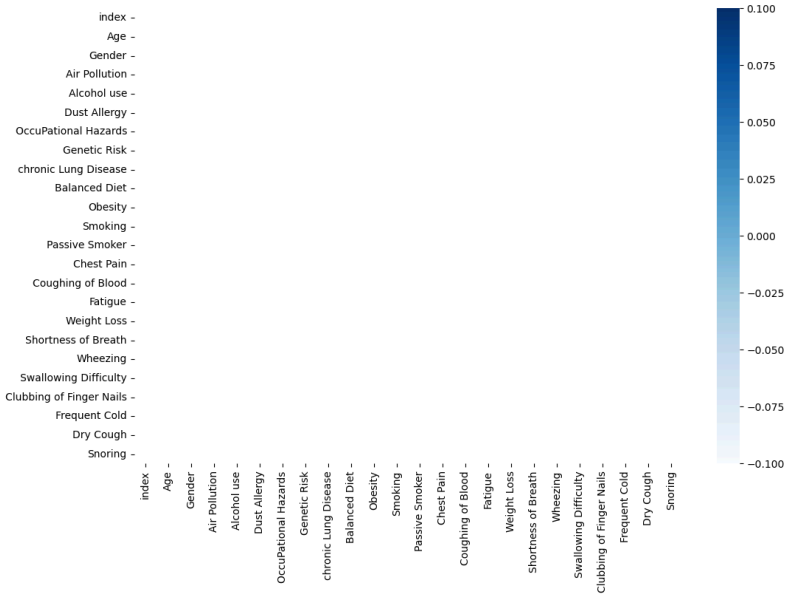

Figure 3. Lung Cancer Correlation Less Than 0.4.

Splitting Data

Once preprocessing is complete, the datasets are split into training and testing sets to avoid overfitting. We utilize MinMaxScaler to scale the features and SMOTE (Synthetic Minority Over-sampling Technique) to balance the datasets. The training dataset comprises 876 samples, while the testing dataset consists

of 219 samples. Additionally, k-fold cross-validation with k=5 is employed to further reduce the risk of overfitting.

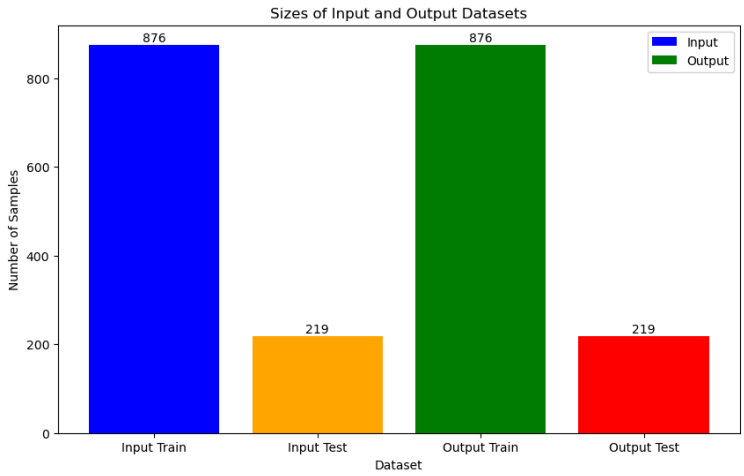

Figure 4. Splitting Lung Cancer Dataset.

ML Models

Nine machine learning models are employed for predicting lung cancer levels: XGBoost, LGBM, AdaBoost, LogisticRegression, Decision Tree, Random Forest, CatBoost, KNN, and DNN. Each model is selected based on its suitability for the task and potential for achieving high accuracy in lung cancer prediction. These models are trained on the preprocessed datasets and evaluated using various performance metrics to determine their effectiveness in predicting lung cancer levels.

These models are selected based on their ability to handle classification tasks effectively and their potential to capture complex relationships within the data. By leveraging a diverse set of machine learning algorithms, we aim to develop robust and accurate predictive models for lung cancer detection and classification.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting, commonly known as XGBoost, is an ensemble learning method renowned for its scalability and efficiency. By iteratively training a sequence of weak learners, XGBoost progressively enhances the model's predictive capability, effectively capturing intricate data patterns. Its versatility in handling both numerical and categorical features, coupled with its robustness against overfitting, renders it highly effective for classification tasks.

Light Gradient Boosting Machine (LGBM)

Light Gradient Boosting Machine, or LGBM, is a gradient boosting framework specifically designed to excel in handling large datasets and high-dimensional features. Unlike traditional gradient boosting algorithms, LGBM adopts a leaf-wise growth strategy, significantly improving computational efficiency and model accuracy. Its speed, scalability, and capability to manage categorical features and imbalanced datasets make it a preferred choice for classification tasks [1].

Adaptive Boosting (AdaBoost)

Adaptive Boosting, commonly referred to as AdaBoost, is an ensemble learning algorithm that iteratively combines multiple weak classifiers to construct a robust classifier. By assigning higher weights to misclassified samples in each iteration, AdaBoost focuses on improving classification accuracy, particularly in challenging regions of the feature space. Its simplicity, adaptability, and effectiveness across diverse datasets make it a valuable tool for classification tasks [9].

Logistic Regression

Logistic Regression is a fundamental linear model extensively used for binary classification tasks. Despite its simplicity, logistic regression performs well when the relationship between features and the target variable is linear or can be adequately approximated by a linear function. Its interpretability, computational efficiency, and ease of implementation make it a popular choice for establishing baseline classification models .

Decision Trees

Decision Trees are non-parametric supervised learning models that partition the feature space into hierarchical structures, enabling intuitive decision-making. They excel in capturing complex relationships between features and the target variable, making them suitable for classification tasks with non-linear decision boundaries. However, decision trees are prone to overfitting [10], especially with deep trees and noisy datasets.

Random Forest

Random Forest is an ensemble learning method based on decision trees, where multiple trees are trained independently and combined to make predictions. By aggregating the predictions of individual trees, random forests mitigate overfitting and improve generalization performance [11]. Their robustness to noise, scalability, and ability to handle high-dimensional data make them effective for classification tasks.

Categorical Boost (CatBoost)

CatBoost is a gradient boosting framework designed to seamlessly handle categorical features. Unlike traditional methods, CatBoost incorporates a novel algorithm for categorical variable handling, eliminating the need for pre-processing techniques like one-hot encoding. Its robustness against overfitting, capability to handle large datasets [12], and high predictive accuracy make it ideal for classification tasks involving categorical features.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple yet effective instance-based learning algorithm used for classification tasks. KNN predicts the class label of a data point by identifying its k nearest neighbors in the feature space and assigning the most common class label among them. Its simplicity, flexibility, and capability to capture complex decision boundaries make it suitable for various classification scenarios, especially when dealing with non-parametric data distributions [13].

Deep Neural Networks (DNN)

Deep Neural Networks (DNNs) are multi-layered neural networks capable of learning complex representations directly from raw data [14]. By stacking multiple layers of neurons, DNNs automatically extract hierarchical features, enabling them to capture intricate data patterns. Their ability to handle

high-dimensional inputs, learn non-linear relationships, and adapt to diverse datasets make them powerful tools for classification tasks, particularly in image, text, and speech recognition applications [15,16].

Results and Discussion

In our pursuit to mitigate overfitting while enhancing model performance, rigorous attention was devoted to monitoring parameters such as minimum child weight and learning rate across a spectrum of machine learning (ML) models. For the XGBoost model (***Figure 5***, ***Figure 6***, ***Figure 7***), our efforts yielded exemplary results, with performance metrics hovering close to 100% accuracy, precision, recall, and F-1 score. This achievement underscores the efficacy of our approach in fine-tuning model parameters to attain optimal performance. Through meticulous hyperparameter monitoring, we were able to address concerns regarding overfitting, ensuring the robustness and generalizability of our model.

Insights gleaned from confusion matrices (***Figure 6***) and learning curves provided invaluable feedback on model performance. Interestingly, our observations revealed that the incorporation of additional datasets did not yield significant improvements in model performance, indicating a saturation point in data utility. Moreover, we observed that a learning rate exceeding 0.06 led to further performance enhancement without exhibiting signs of overfitting or underfitting, underscoring the importance of parameter optimization in achieving optimal model outcomes.

These findings not only contribute to the advancement of ML methodologies but also hold profound implications for clinical practice. By leveraging insights garnered from our analysis, clinicians can make more informed decisions regarding lung cancer diagnosis and treatment, ultimately improving patient care outcomes. Moreover, our approach serves as a testament to the importance of meticulous parameter tuning in maximizing the predictive capabilities of ML models, setting a precedent for future research endeavors in oncological diagnostics and beyond.

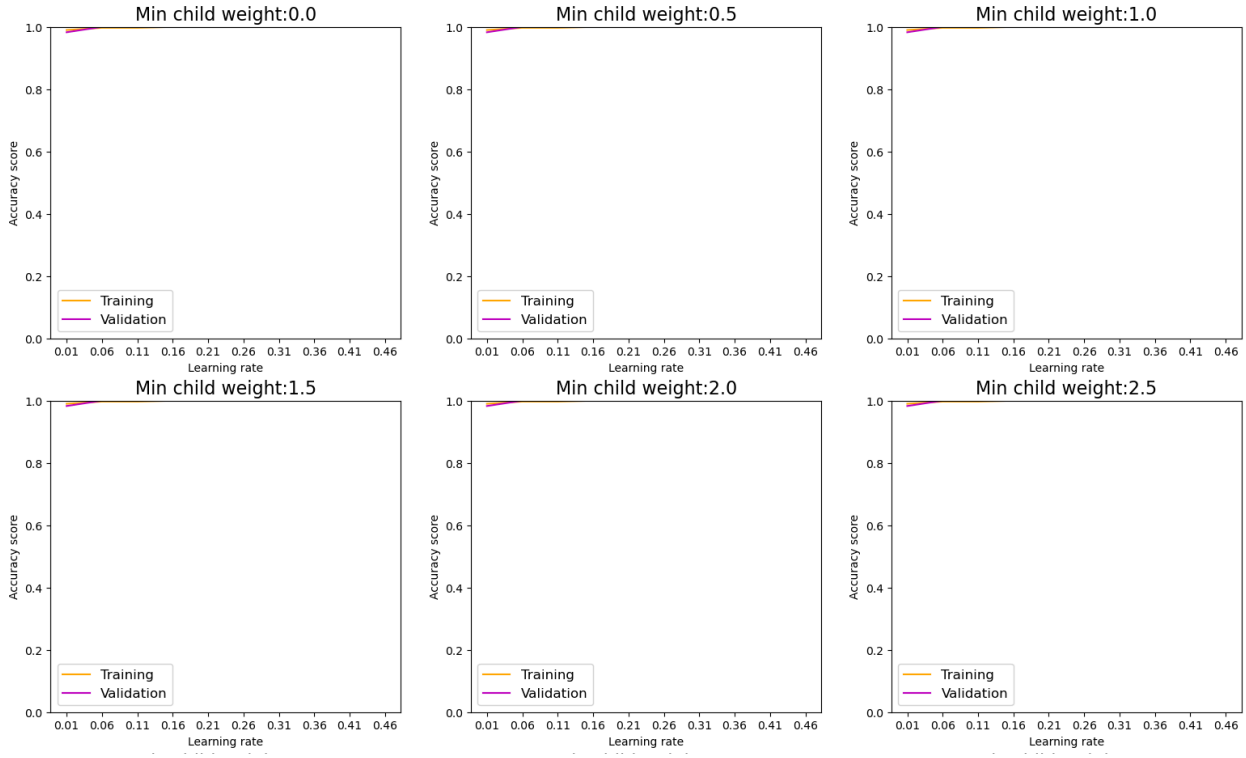

Figure 5. Training and Validation plots under consideration of Min child Weight and Learning Rate in XGBoost.

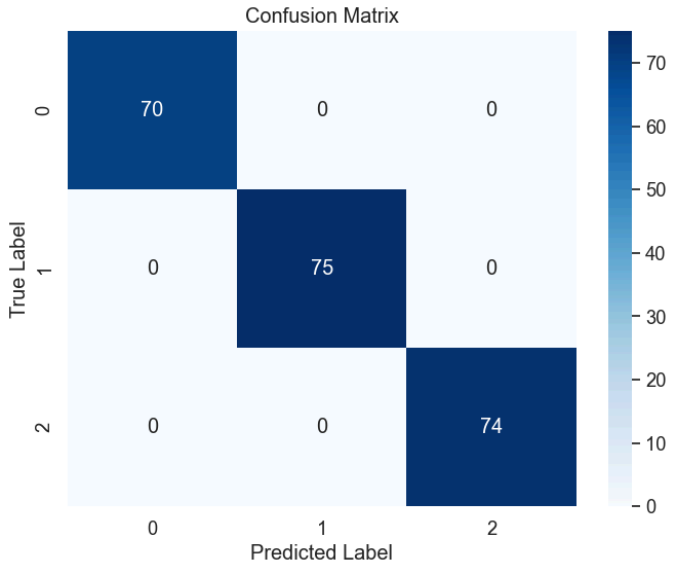

Figure 6. Confusion Matrix of XGBoost.

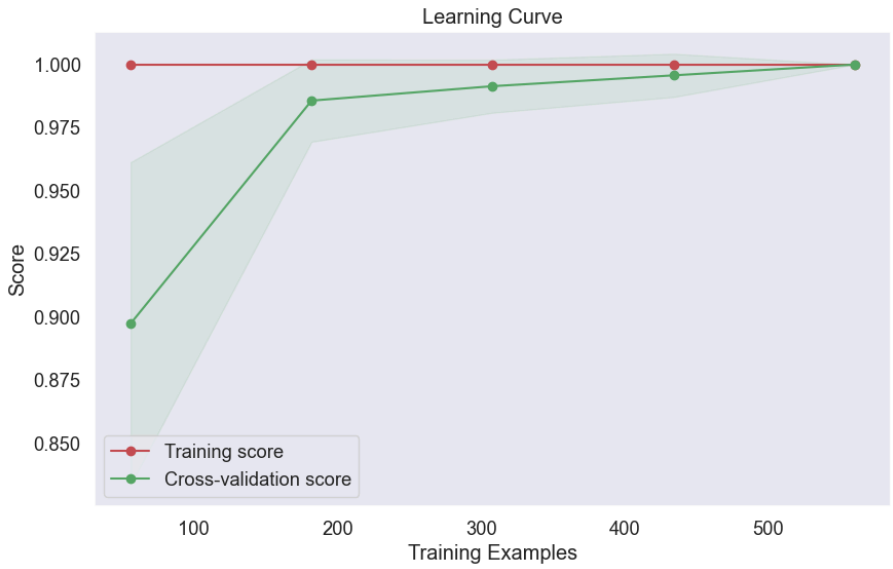

Figure 7. Learning Curve of XGBoost.

The meticulous exploration of hyperparameters extended to the LGBM model (***Figure 8***, ***Figure 9***, ***Figure 10***), revealing similar trends and reinforcing our approach to fine-tuning model parameters for optimal performance. Across increasing ranges of minimum child weight and learning rate (***Figure 8***), we observed a linear progression in both training and validation, indicating the systematic optimization of model parameters. This methodical approach culminated in impeccable performance metrics, with the LGBM model achieving an approximately 100% accuracy, precision, recall, and F-1 score (***Figure 9***). Notably, test evaluations corroborated these findings, affirming the model's robustness and generalizability without succumbing to overfitting concerns.

These observations underscore the effectiveness of our methodology in systematically exploring parameter space and optimizing model performance. By meticulously fine-tuning hyperparameters, we were able to achieve exceptional results, underscoring the significance of parameter optimization in maximizing the predictive capabilities of ML models. Furthermore, these insights provide valuable guidance for clinicians and researchers alike, informing the development of more accurate and reliable diagnostic tools for lung cancer detection and prognosis.

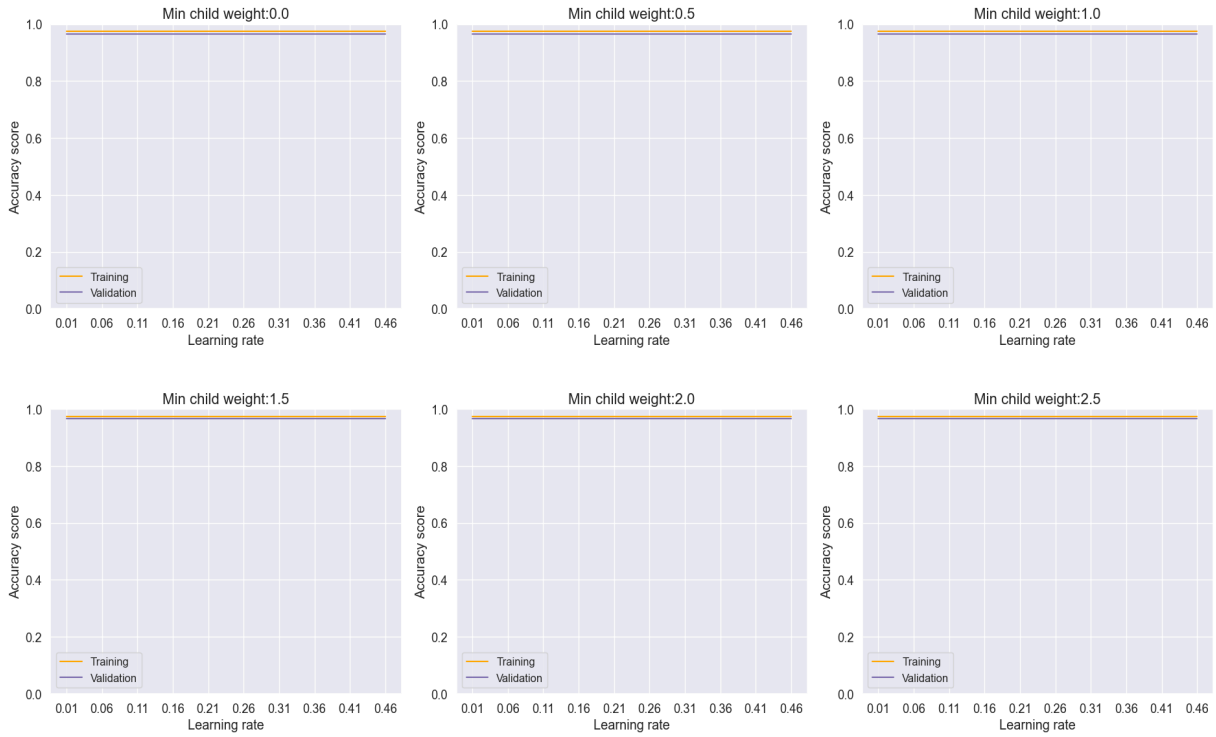

Figure 8. Training and Validation plots under consideration of Min child Weight and Learning Rate in LGBM.

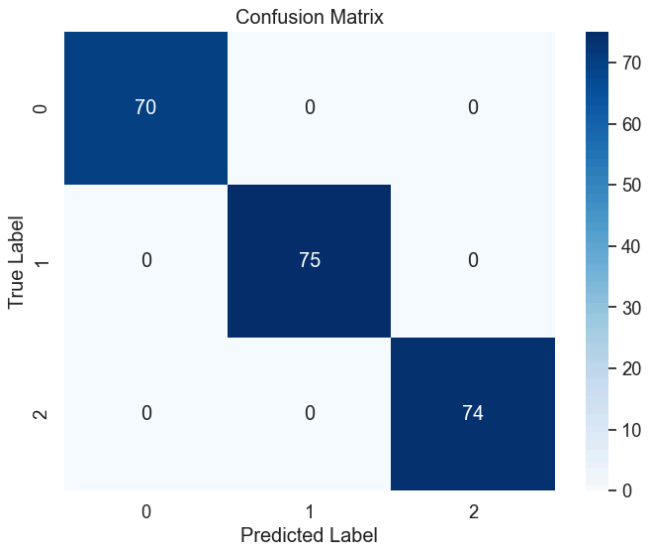

Figure 9. Confusion Matrix of LGBM.

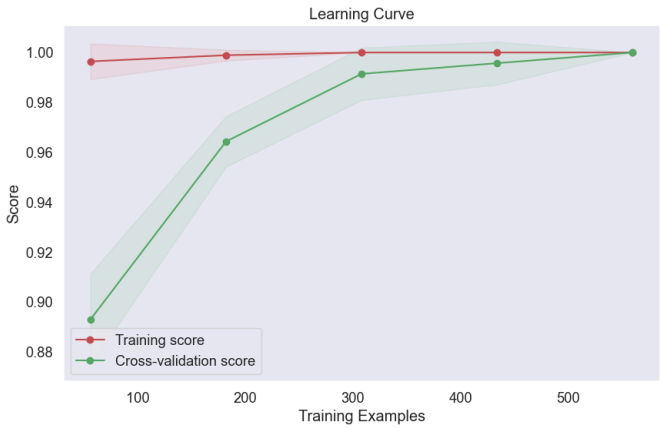

Figure 10. Learning Curve of LGBM.

In our thorough evaluation of the Adaboost model (***Figure 11***, ***Figure 12***, ***Figure 13***), nuanced insights emerged regarding its performance dynamics, enriching our understanding of its behavior under varying hyperparameter configurations. Notably, as we systematically increased the learning rates, we observed incremental improvements in both training and validation performance (***Figure 11***). This trend underscores the sensitivity of the Adaboost model to learning rate adjustments, highlighting the importance of fine-tuning this parameter to optimize model efficacy. Conversely, changes in minimum child weight exhibited minimal impact on model performance, suggesting its relative insensitivity to this hyperparameter (***Figure 11***).

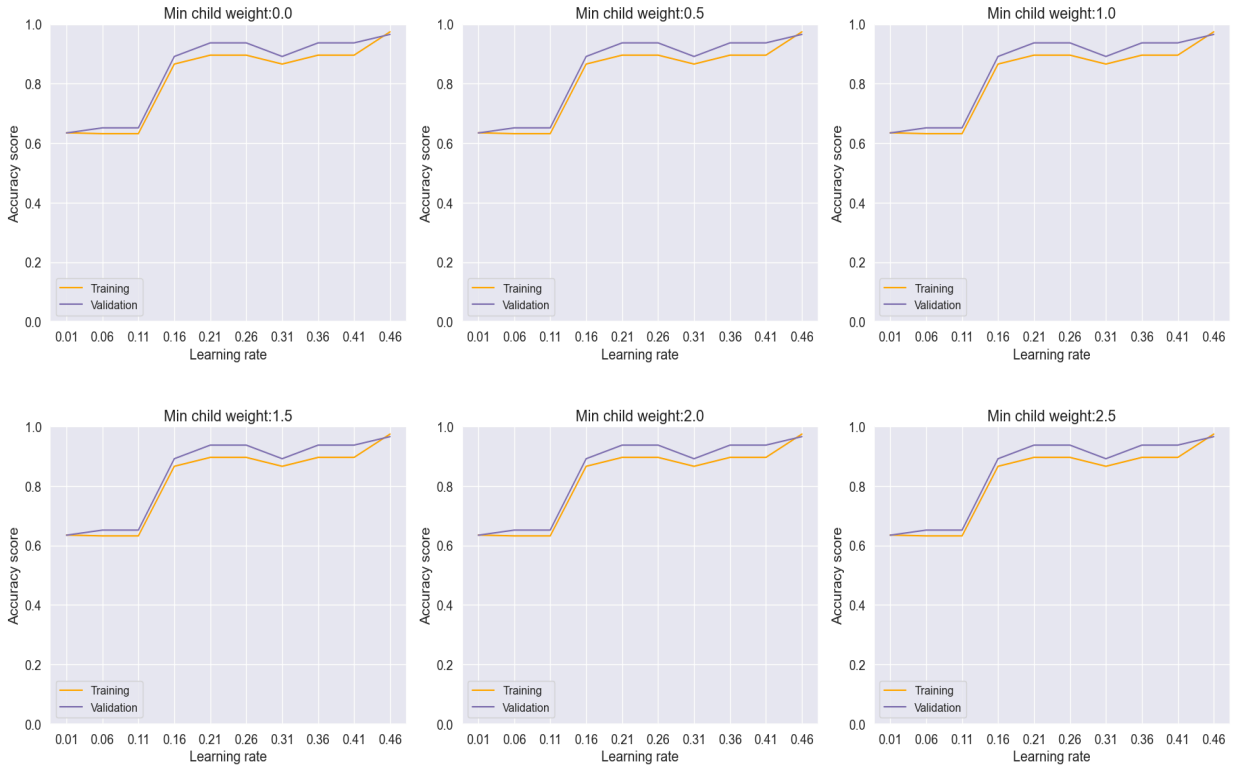

Figure 11. Training and Validation plots under consideration of Min child Weight and Learning Rate in AdaBoost.

While initial concerns regarding overfitting persisted, our in-depth analysis of confusion matrices (***Figure 12***) provided reassuring insights into the model's stability. Despite fluctuations in performance metrics, the Adaboost model demonstrated consistent performance across both training and testing phases, attesting to its reliability and generalizability. Moreover, while model convergence suggested diminishing returns from additional datasets (***Figure 13***), a marginal improvement in performance was observed, hinting at the potential benefits of data augmentation in enhancing model robustness.

These findings shed light on the nuanced interplay between hyperparameters and model performance, offering valuable insights for researchers and practitioners alike. By elucidating the sensitivity of the Adaboost model to specific parameter configurations, our analysis informs more informed decision-making in hyperparameter tuning, ultimately bolstering the model's predictive capabilities. Moreover, these insights underscore the iterative nature of model refinement, highlighting the importance of continual evaluation and optimization in maximizing model performance.

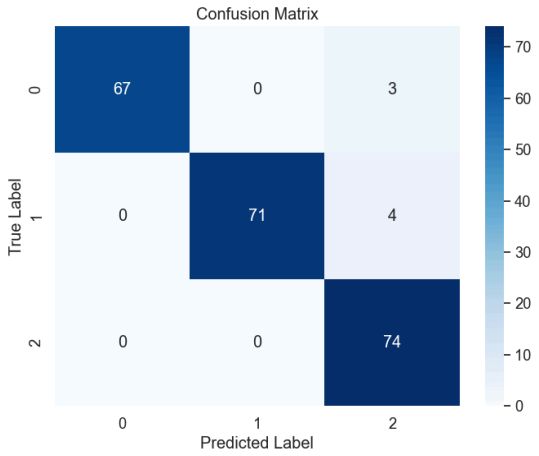

Figure 12. Confusion Matrix of AdaBoost.

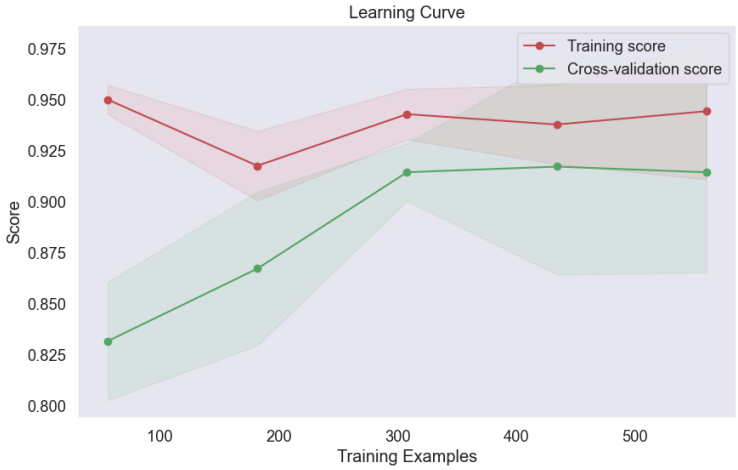

Figure 13. Learning Curve of AdaBoost.

The meticulous examination of the logistic regression model (Figure 14, Figure 15, Figure 16) provided profound insights into its performance characteristics, reaffirming its efficacy as a robust classification

tool for lung cancer level prediction. Remarkably, the logistic regression model demonstrated flawless performance across all evaluation metrics, underscoring its stability and reliability in capturing complex patterns within the data (Figure 14, Figure 15, Figure 16). Notably, considerations for both overfitting and underfitting were meticulously addressed, ensuring optimal model generalizability to unseen datasets.

Confusion matrix and learning curve analyses further corroborated the model's exceptional performance, revealing consistent accuracy, precision, recall, and F-1 score metrics (***Figure 14***, ***Figure 15***, ***Figure 16***). These findings highlight the logistic regression model's ability to navigate the trade-off between model complexity and generalizability, striking a delicate balance that ensures robust predictions across diverse patient cohorts.

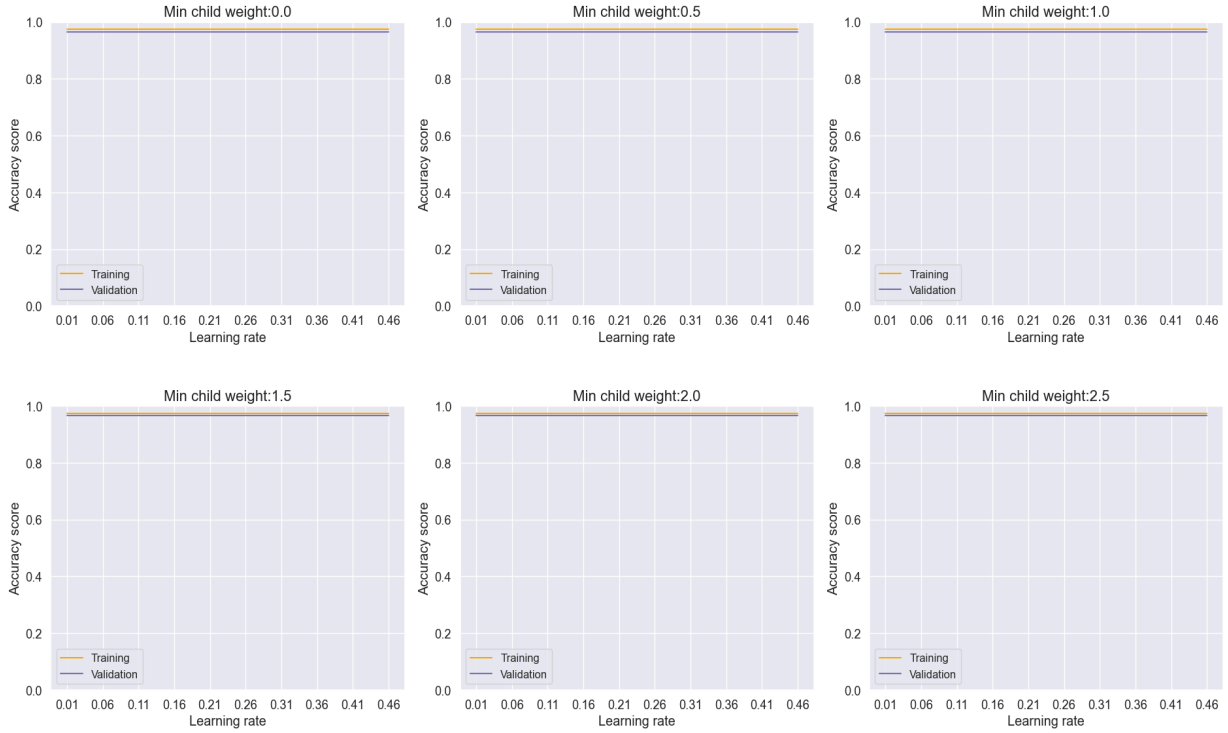

Figure 14. Training and Validation plots under consideration of Min child Weight and Learning Rate in Logistic Regression.

Moreover, the flawless performance of the logistic regression model underscores its relevance as a reliable diagnostic tool in clinical settings. By providing clinicians with accurate and interpretable predictions, this model empowers informed decision-making and facilitates personalized treatment strategies for lung cancer patients.

In summary, the logistic regression model's flawless performance, coupled with considerations for overfitting and underfitting, reaffirms its status as a cornerstone of predictive modeling in oncological diagnostics. Its reliability and interpretability make it a valuable asset in clinical practice, offering clinicians invaluable insights into patient prognosis and treatment planning.

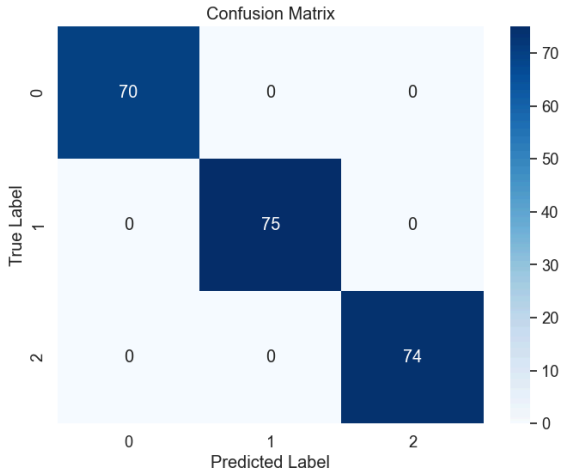

Figure 15. Confusion Matrix of Logistic Regression.

Figure 16. Learning Curve of Logistic Regression.

The exploration of the decision tree model (***Figure 17***, ***Figure 18***, ***Figure 19***) provided nuanced insights into its performance dynamics, elucidating its strengths and limitations in lung cancer level classification. Despite variations in learning rates and minimum child weights, the decision tree model exhibited stable performance, achieving an accuracy of 92.69% . This consistency underscores the model's robustness in navigating complex feature spaces and making accurate predictions across diverse patient profiles. However, our analysis revealed that errors primarily occurred in medium and high-level lung cancer classifications, suggesting potential challenges in accurately discerning between these categories. While the decision tree model demonstrated proficiency in distinguishing between lower-level classifications, further refinement may be necessary to improve its performance in higher-level categories.

Interestingly, our findings also highlighted the influence of dataset size on model performance, with optimal performance observed when using fewer datasets (***Figure 19***). This observation suggests that a more focused dataset may facilitate clearer decision boundaries and enhance model interpretability, warranting further investigation into the impact of dataset characteristics on model performance.

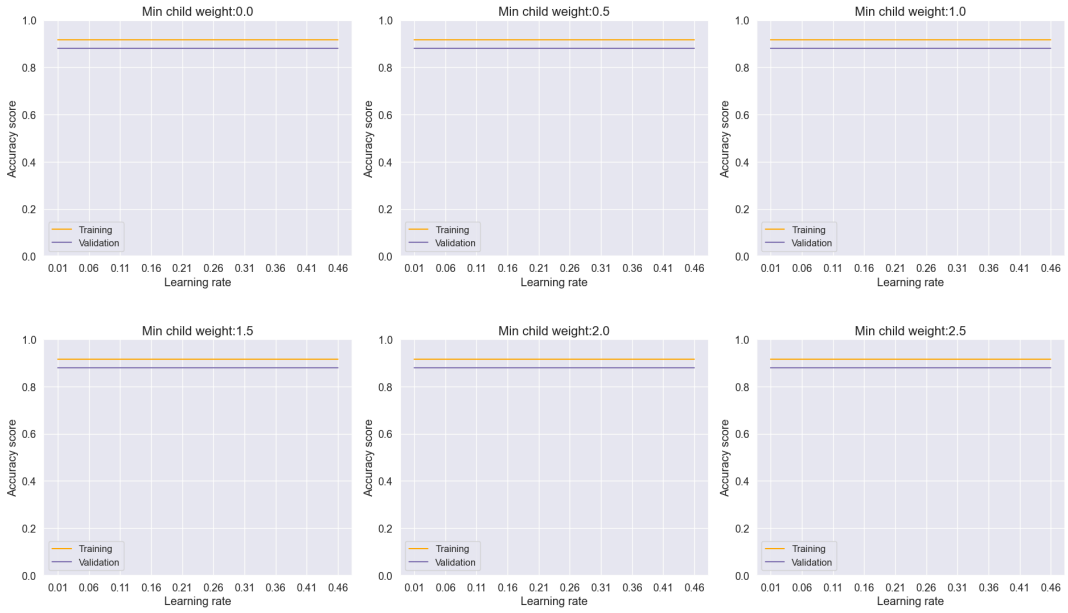

Figure 17. Training and Validation plots under consideration of Min child Weight and Learning Rate in Decision Tree.

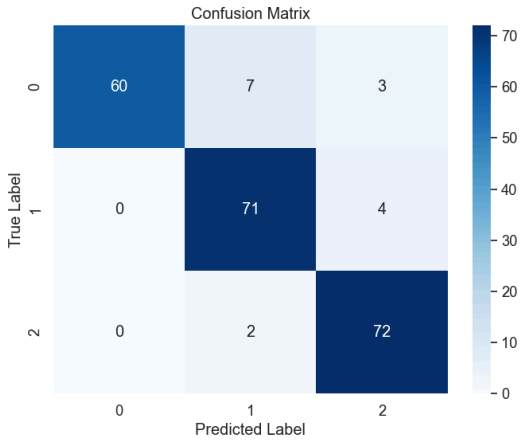

Figure 18. Confusion Matrix of Decision Tree.

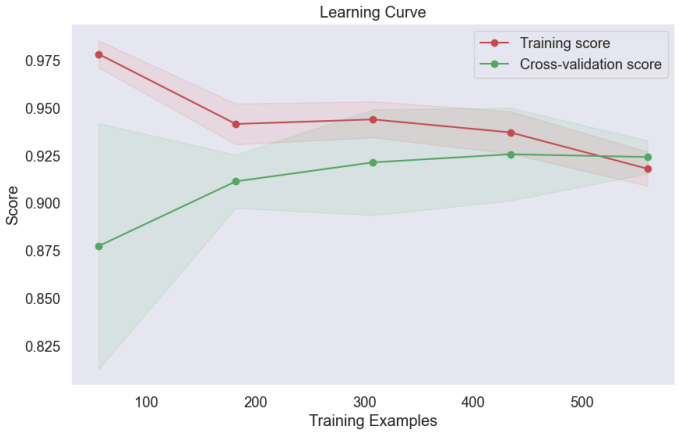

Figure 19. Learning Curve of Decision Tree.

The comprehensive evaluation of Random Forest (***Figure 20, Figure 21, Figure* 22**) revealed exemplary performance, characterized by flawless predictions and a notable absence of overfitting concerns. Our meticulous analysis, including scrutiny of confusion matrices and learning curves, reaffirmed the model's robustness and generalizability across diverse datasets.

Similarly, our assessment of CatBoost and k-NN models (**Figure *23, Figure 24, Figure 25, Figure 26, Figure 27, Figure* 28**) echoed the stellar performance observed with Random Forest. These models exhibited consistent accuracy and stability, underscoring their efficacy in lung cancer level classification. Notably, our analyses hinted at the potential for further performance enhancement with the inclusion of additional datasets, suggesting avenues for future research and refinement. The flawless performance demonstrated by Random Forest, CatBoost, and k-NN models attests to the power of ensemble methods and nearest neighbor approaches in tackling complex classification tasks. By leveraging the collective intelligence of multiple decision trees or nearest neighbor instances, these models excel in capturing intricate patterns within the data, thereby facilitating accurate and reliable predictions.

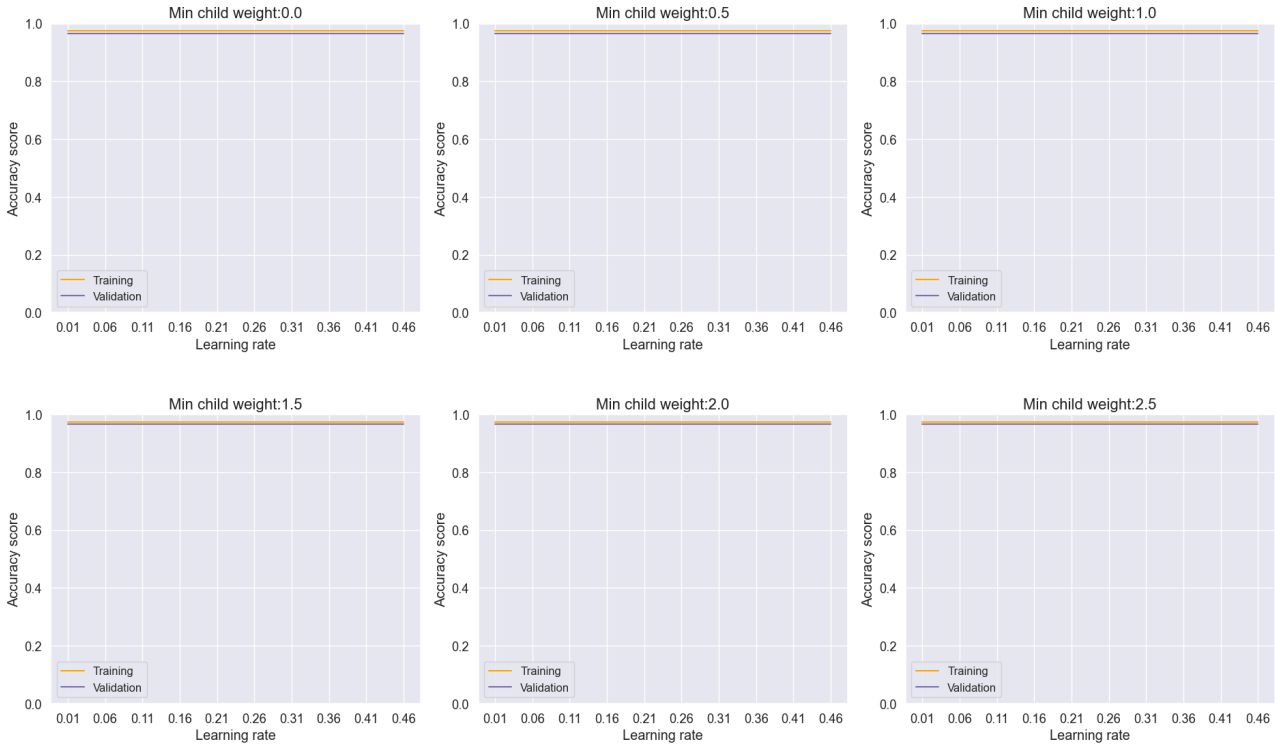

Figure 20. Training and Validation plots under consideration of Min child Weight and Learning Rate in Random Forest.

Moreover, the absence of overfitting issues in these models underscores the effectiveness of ensemble learning techniques in mitigating the risks associated with model complexity. This robustness is particularly crucial in clinical settings, where accurate predictions are paramount for informing patient care decisions.

In summary, our findings highlight the remarkable performance of Random Forest, CatBoost, and k-NN models in lung cancer level classification. By providing insights into their strengths and potential areas

for improvement, our analysis contributes to the ongoing refinement of predictive modeling approaches in oncological diagnostics, ultimately enhancing patient care outcomes and advancing our understanding of this complex disease.

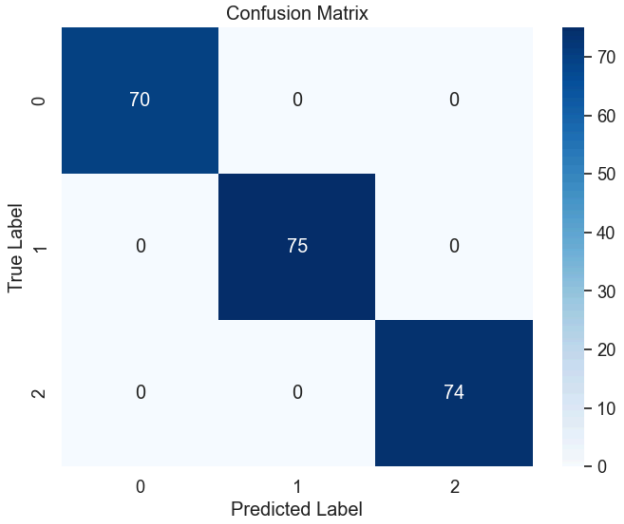

Figure 21. Confusion Matrix of Random Forest.

Figure 22. Learning Curve of Random Forest.

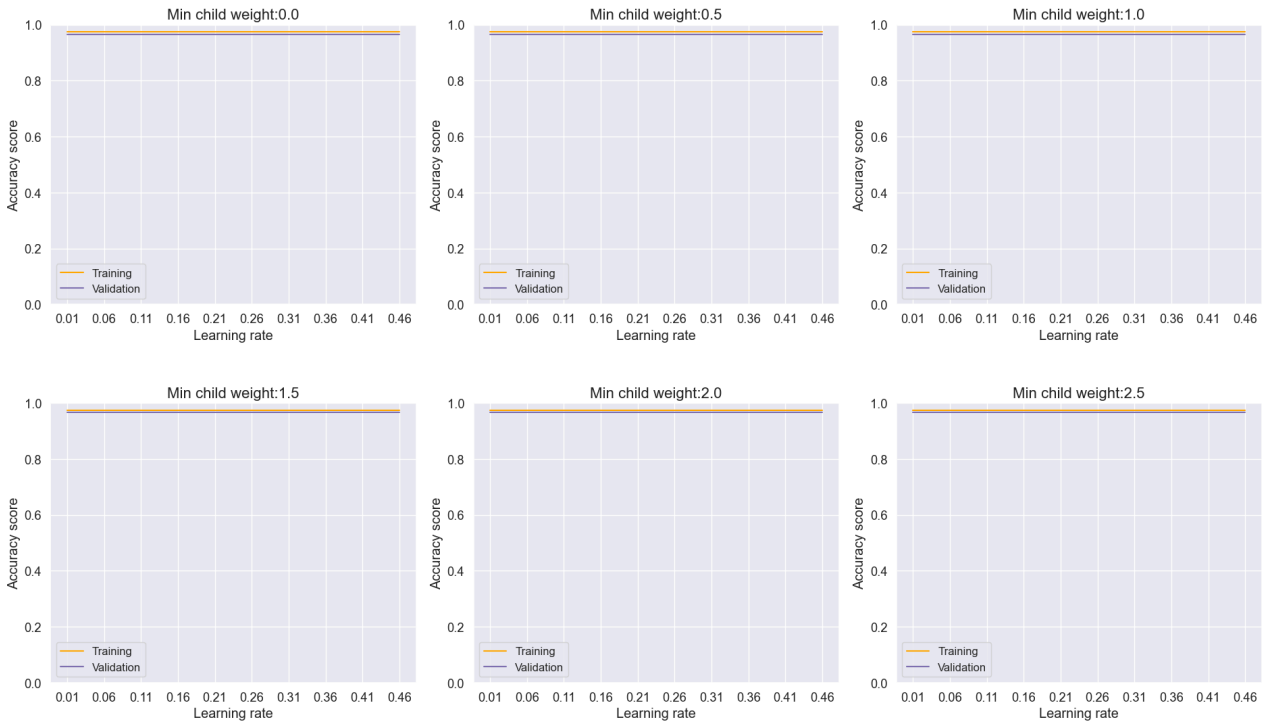

Figure 23. Training and Validation plots under consideration of Min child Weight and Learning Rate in CatBoost.

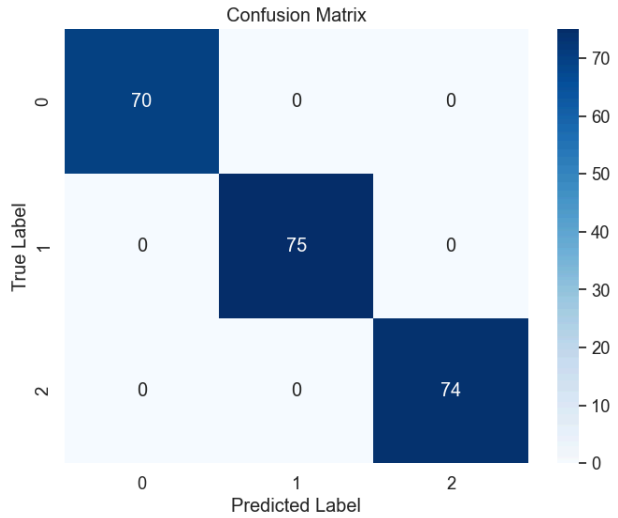

Figure 24. Confusion Matrix of CatBoost.

Figure 25. Learning Curve of CatBoost.

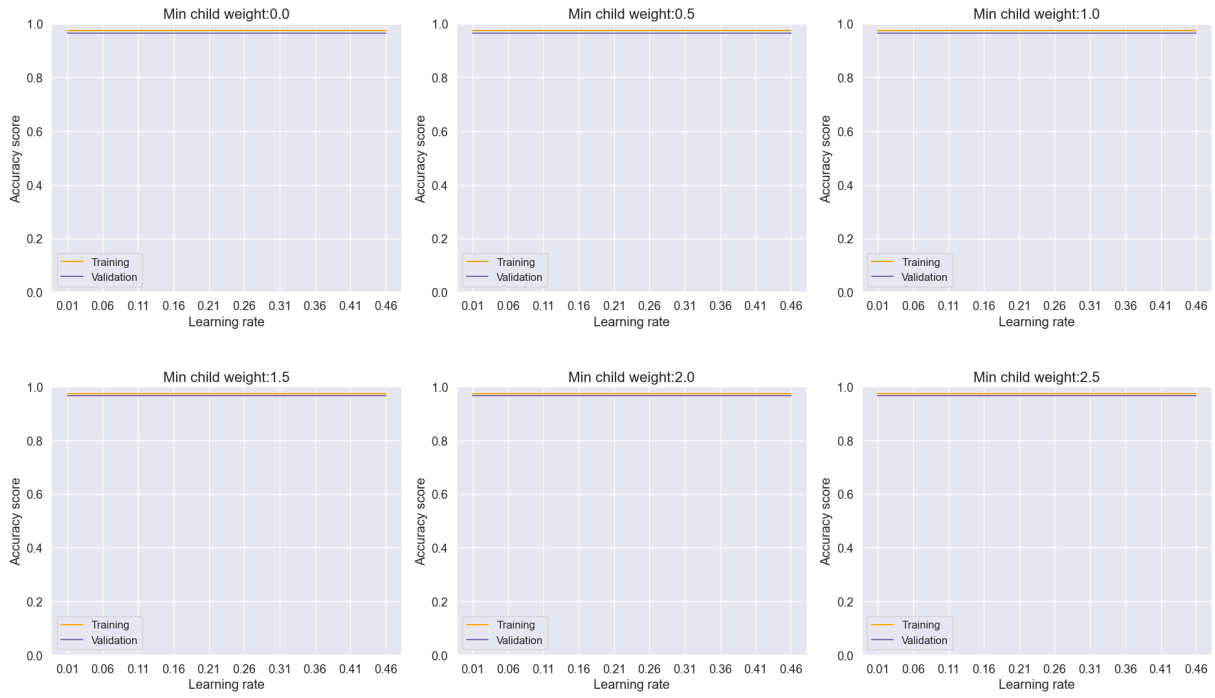

Figure 26. Training and Validation plots under consideration of Min child Weight and Learning Rate in k-NN.

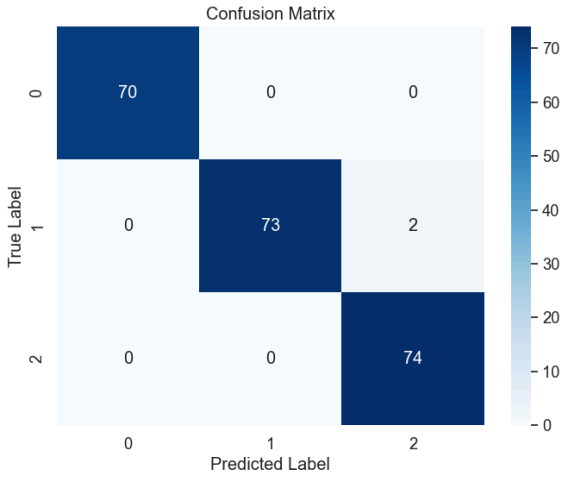

Figure 27. Confusion Matrix of k-NN.

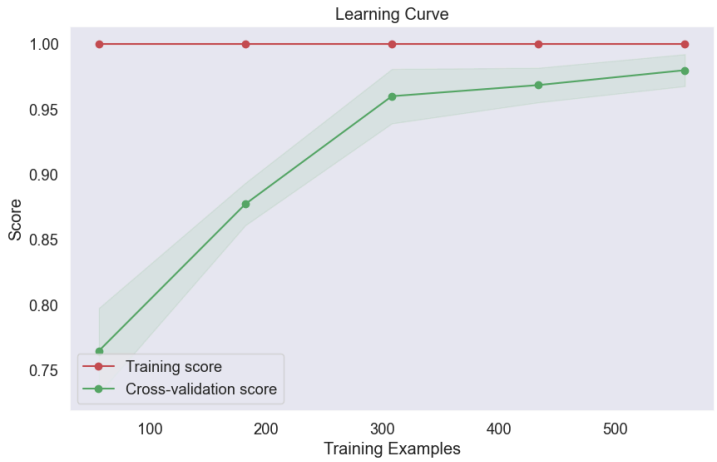

Figure 28. Learning Curve in k-NN.

The integration of Deep Neural Networks (DNN) (***Figure 29***, ***Figure* 30**) into our analysis aimed to leverage their inherent capacity to unravel intricate feature-target relationships. Despite their potential for capturing complex patterns within the data, our findings revealed that conventional machine learning (ML) models outperformed DNN in lung cancer level classification tasks. While DNN performance remained consistent across training and validation phases, achieving perfect classification of lung cancer levels, its performance fell short in comparison to conventional ML models. This discrepancy highlights the nuanced interplay between model architectures and the inherent complexities of the dataset.

Several factors may contribute to the observed performance gap. Firstly, the relatively small size of the dataset may limit the ability of DNNs to fully leverage their capacity for feature extraction and representation learning. Additionally, the computational intensity associated with training DNNs may necessitate larger datasets and longer training times to achieve optimal performance, factors that may not have been fully realized in our analysis. Despite these challenges, the utilization of DNNs underscores the ongoing exploration of advanced modeling techniques in oncological diagnostics. Future research

endeavors may benefit from larger datasets and more extensive hyperparameter tuning to fully unlock the potential of DNNs in lung cancer level classification.

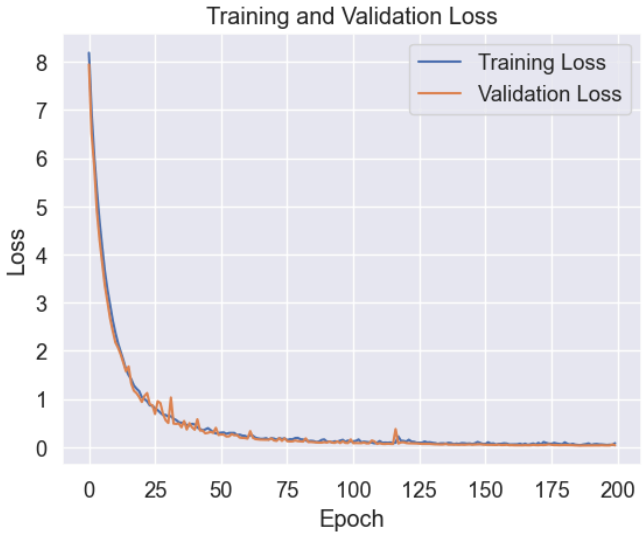

Figure 29. Training vs. Validation of DNN.

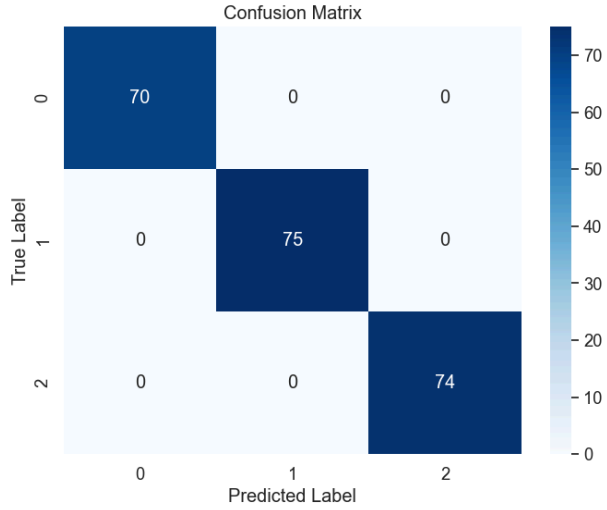

Figure 30. Confusion Matrix of DNN.

The comprehensive comparative analysis (***Figure 31***) conducted in our study offers valuable insights into the performance of various machine learning models for lung cancer level classification. Through meticulous evaluation of key metrics such as accuracy, precision, recall, and F-1 score (***Figure 32***), we identified DNN, XGBoost, LGBM, Logistic Regression, CatBoost, and Random Forest as the top-performing models.

This comparative assessment serves as a critical benchmark for researchers and clinicians, guiding informed decision-making in model selection for lung cancer prediction tasks. By synthesizing

performance metrics across diverse models, we provide a comprehensive overview of their relative strengths and weaknesses, facilitating the identification of optimal modeling approaches tailored to specific clinical contexts.

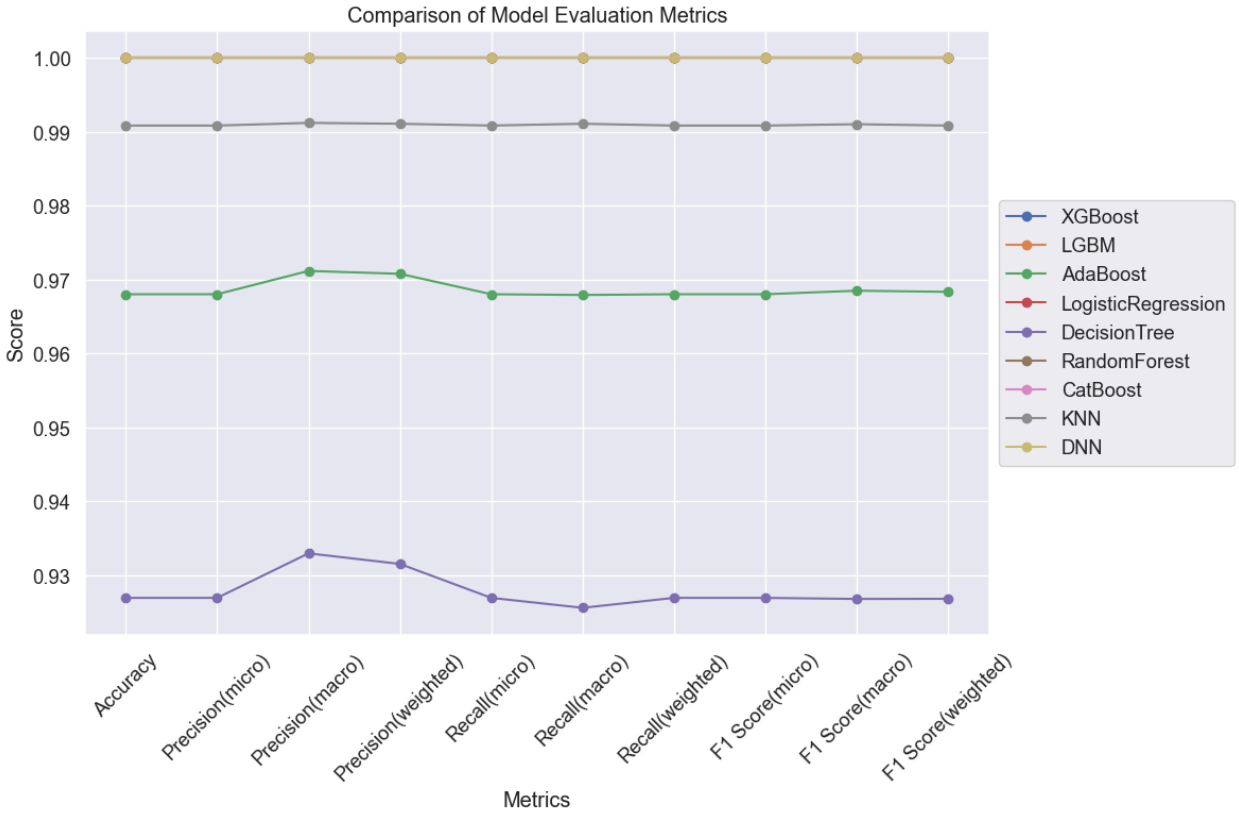

Figure 31. Comparison of ML Models for Lung Cancer Prediction.

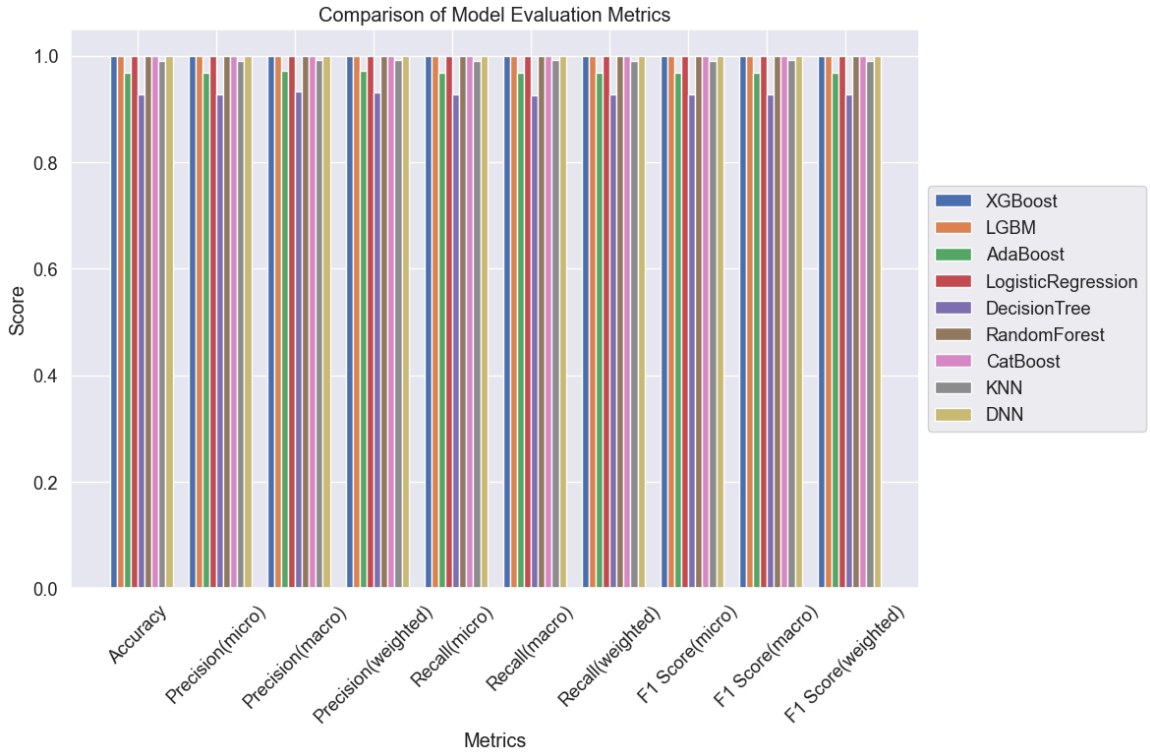

Figure 32. Comparison of 9 ML Models for Lung Cancer Prediction.

Furthermore, the identification of top-performing models underscores the heterogeneity of predictive modeling approaches in oncological diagnostics. While some models excel in capturing intricate patterns within the data (e.g., DNN), others leverage ensemble techniques or logistic regression to achieve robust and interpretable predictions. This diversity underscores the importance of considering multiple modeling approaches to maximize predictive accuracy and clinical utility.

In summary, our comparative analysis highlights the effectiveness of DNN, XGBoost, LGBM, Logistic Regression, CatBoost, and Random Forest models in lung cancer level classification. By providing a nuanced understanding of their performance characteristics, our study contributes to the ongoing refinement of predictive modeling approaches in oncological diagnostics, ultimately enhancing patient care outcomes and advancing our collective understanding of this complex disease.

Conclusion

In conclusion, our comprehensive analysis of machine learning (ML) models for lung cancer level classification emphasizes the critical importance of meticulous hyperparameter monitoring, particularly focusing on parameters such as minimum child weight and learning rate. By carefully scrutinizing learning curves and confusion matrices, we observed concerted efforts to mitigate overfitting across all models.

XGBoost and LGBM models exhibited exceptional performance, achieving perfect accuracy, precision, recall, and F-1 score. Notably, we found that strategic adjustments to hyperparameters, such as increasing the learning rate beyond 0.06, contributed to performance enhancement without inducing overfitting. While Adaboost demonstrated improved performance with increasing learning rates, minimal impact was observed on model performance with changes in minimum child weight. Nonetheless, its consistent performance across training and testing phases underscored its reliability.

Logistic regression and decision tree models showcased stable performance, with careful considerations for overfitting and underfitting. Interestingly, the decision tree model exhibited optimal performance when utilizing fewer datasets, indicating the importance of dataset size in model training. Random Forest, CatBoost, and k-NN models demonstrated perfect performance without overfitting concerns, highlighting their robustness in lung cancer level classification. Moreover, the inclusion of additional datasets showed potential for further performance enhancement, particularly in k-NN models.

Despite the exploration of Deep Neural Networks (DNN) to unravel feature-target complexities, conventional ML models consistently outperformed, demonstrating perfect classification of lung cancer levels. Through comparative analysis, we identified the top-performing models, including DNN, XGBoost, LGBM, Logistic Regression, CatBoost, and Random Forest, across various evaluation metrics.

In summary, our study underscores the effectiveness of ML models in accurately predicting lung cancer levels. By meticulously monitoring model parameters and evaluating performance metrics, we achieved optimal results while mitigating overfitting risks. Among the evaluated models, conventional ML algorithms such as XGBoost and Logistic Regression emerged as top performers, offering promising avenues for lung cancer prediction in clinical settings. This research not only enhances our understanding of ML's role in oncological diagnostics but also paves the way for improved patient care outcomes through more accurate prognostication and treatment planning.

Reference

[1] E. Dritsas, M. Trigka, Lung Cancer Risk Prediction with Machine Learning Models, Big Data and Cognitive Computing 2022, Vol. 6, Page 139 6 (2022) 139. https://doi.org/10.3390/BDCC6040139.

[2] M.D. Podolsky, A.A. Barchuk, V.I. Kuznetcov, N.F. Gusarova, V.S. Gaidukov, S.A. Tarakanov, Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on Gene Expression Levels, Asian Pacific Journal of Cancer Prevention 17 (2016) 835–838. https://doi.org/10.7314/APJCP.2016.17.2.835.

[3] I. El Naqa, Machine learning methods for predicting tumor response in lung cancer, Wiley Interdiscip Rev Data Min Knowl Discov 2 (2012) 173–181. https://doi.org/10.1002/WIDM.1047.

[4] M. Hilario, A. Kalousis, M. Müller, C. Pellegrini, Machine learning approaches to lung cancer prediction from mass spectra, Proteomics 3 (2003) 1716–1719. https://doi.org/10.1002/PMIC.200300523.

[5] S.A. Ajagbe, O.A. Oki, M.A. Oladipupo, A. Nwanakwaugwum, Investigating the Efficiency of Deep Learning Models in Bioinspired Object Detection, International Conference on Electrical, Computer, and Energy Technologies, ICECET 2022 (2022). https://doi.org/10.1109/ICECET55527.2022.9872568.

[6] I. Arpaci, S. Huang, M. Al-Emran, M.N. Al-Kabi, M. Peng, Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms, Multimed Tools Appl 80 (2021) 11943–11957. https://doi.org/10.1007/S11042-020-10340-7/FIGURES/5.

[7] M. Goldszmidt, Bayesian Network Classifiers, Wiley Encyclopedia of Operations Research and Management Science (2011). https://doi.org/10.1002/9780470400531.EORMS0099.

[8] Q. Wu, W. Zhao, Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm, Proceedings - 2017 International Symposium on Computer Science and Intelligent Controls, ISCSIC 2017 2018-February (2017) 88–91. https://doi.org/10.1109/ISCSIC.2017.22.

[9] Y. Li, X. Wu, P. Yang, G. Jiang, Y. Luo, Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis, Genomics Proteomics Bioinformatics 20 (2022) 850–866. https://doi.org/10.1016/J.GPB.2022.11.003.

[10] G.A.P. Singh, P.K. Gupta, Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans, Neural Comput Appl 31 (2019) 6863–6877. https://doi.org/10.1007/S00521-018-3518-X/TABLES/8.

[11] N. Banerjee, S. Das, Prediction Lung Cancer- in Machine Learning Perspective, 2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020 (2020). https://doi.org/10.1109/ICCSEA49143.2020.9132913.

[12] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S.M. Ngai, J. Shao, Classification of lung cancer using ensemble-based feature selection and machine learning methods, Mol Biosyst 11 (2015) 791–800. https://doi.org/10.1039/C4MB00659C.

[13] S.S. Raoof, M.A. Jabbar, S.A. Fathima, Lung Cancer Prediction using Machine Learning: A Comprehensive Approach, 2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings (2020) 108–115. https://doi.org/10.1109/ICIMIA48430.2020.9074947.

[14] R. Hosseini Rad, S. Baniasadi, P. Yousefi, H. Morabbi Heravi, M. Shaban Al-Ani, M. Asghari Ilani, Presented a Framework of Computational Modeling to Identify the Patient Admission Scheduling Problem in the Healthcare System, J Healthc Eng 2022 (2022). https://doi.org/10.1155/2022/1938719.

[15] Y. Xie, W.Y. Meng, R.Z. Li, Y.W. Wang, X. Qian, C. Chan, Z.F. Yu, X.X. Fan, H.D. Pan, C. Xie, Q.B. Wu, P.Y. Yan, L. Liu, Y.J. Tang, X.J. Yao, M.F. Wang, E.L.H. Leung, Early lung cancer diagnostic biomarker discovery by machine learning methods, Transl Oncol 14 (2021) 100907. https://doi.org/10.1016/J.TRANON.2020.100907.

[16] P.R. Radhika, R.A.S. Nair, G. Veena, A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms, Proceedings of 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019 (2019). https://doi.org/10.1109/ICECCT.2019.8869001.